

Adaptive gene regulatory networks

Franck Stauffer¹ and Johannes Berg²

¹*Institut für Theoretische Physik, Universität zu Köln
Zùlpicher Straße 77, 50937 Köln, Germany*

²*Physikalisches Institut Albert-Ludwigs-Universität Freiburg
Herrmann Herder-Str. 3, 79104 Freiburg, Germany*

(Dated: February 17, 2009)

Regulatory interactions between genes show a large amount of cross-species variability, even when the underlying functions are conserved: There are many ways to achieve the same function. Here we investigate the ability of regulatory networks to reproduce given expression levels within a simple model of gene regulation. We find an exponentially large space of regulatory networks compatible with a given set of expression levels, giving rise to an extensive *entropy of networks*. Typical realisations of regulatory networks are found to share a bias towards symmetric interactions, in line with empirical evidence.

PACS numbers: 87.16.Yc 87.23.Kg 87.18.Sn

The expression of genes is regulated such that the right combinations of gene products are generated at the right time and place of an organism. Key regulators of gene expression are *transcription factors*, proteins which bind to specific sites on DNA and influence the expression of nearby genes. Typically, the expression of a gene is effected by a combination of several transcription factors, and conversely, a transcription factor regulates several genes. Expression levels can thus depend on the entire set of regulatory interaction between transcription factors and their target genes, referred to as a *regulatory network*. These intracellular reaction networks process extracellular information to induce specific gene expression patterns, allowing, for instance, the development of a complex body plan, or responses to external conditions.

Even though regulatory networks are tuned carefully to produce specific expression patterns, there are in general many networks fulfilling a regulatory task. An example is the control of mating type in different yeast species: The same set of genes controlled in *S. cerevisiae* by an activator which is upregulated in a certain state is controlled by a repressor which is downregulated in that state in *C. albicans* [1]. A second prominent example is the development of the anterior patterning in insect embryos, leading to the formation of the insect's head. The gene crucial to this process in the fruit fly *Drosophila*, called bicoid, is absent in many other insects, where a combination of different genes take on the same task [2]. Even whole sets of genes which are co-expressed across the entire yeast family can have different regulatory interactions in different species [3]. Source of these changing interactions is a rapid evolutionary turnover of transcription factor binding sites at the level of DNA sequences [4, 5]. This can generate new regulatory interactions. A recent essay on the degeneracy of regulatory networks can be found in [6].

The large number of regulatory networks with a given function (viable networks) is particularly relevant from

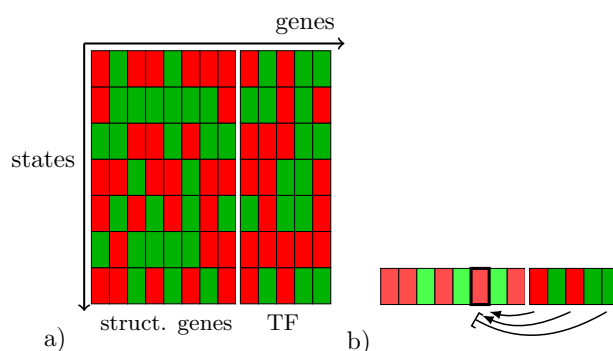


Figure 1: Expression levels and regulatory networks. a) We list genes along the x -axis, and states of the organism along the y -axis. Following established convention in expression analysis, expression levels are colour-coded with high expression levels shown in red (dark), low levels in green (light). b) Regulatory interactions must be compatible with gene expression levels in all states of the organism. The schematic example shows interactions between transcription factors and a single target gene; two enhancing interactions (\rightarrow) with up-regulated transcription factors, and a repressive interaction (\leftarrow) with a downregulated transcription factor lead to the activation of the target gene.

an evolutionary perspective, as neutral evolution gradually explores different viable networks. Viable networks can form a set with intricate geometric features in the space of regulatory network. An analogy is the set of all RNA-sequences which fold into a given secondary structure, which stretches across the entire sequence space [7]. Numerical studies, based on simple models of gene regulation [8] found that the space of viable networks can be crossed in small steps, and that a wide range of new expression patterns can be generated by small changes to different viable networks [9, 10].

These observations call for a statistical approach based on the ensemble of all viable networks, which is the topic of this paper. We consider a model with two classes of

genes: *structural genes* (coding e.g. for enzymes or cellular components), and *transcription factors*. The expression levels of structural genes are prescribed for different states of the organism and are *fixed* for a given state. For instance, when nutrients are available, specific enzymes have to be produced to digest these nutrients. On the other hand, the expression levels of transcription factors, and the regulatory interactions between genes can be *adapted* to meet the expression levels of structural genes. The freedom to alter expression levels of transcription factors turns out to be crucial.

In the following, we develop a simple model based on *quenched* random expression levels of structural genes, and *adaptive* regulatory interactions and expression levels of transcription factors. The ensemble of viable networks is characterized by the microcanonical partition function

$$Z \equiv e^S = \frac{\text{Tr}_{\mathbf{J}, \xi_t} I(\mathbf{J}, \xi)}{\text{Tr}_{\mathbf{J}, \xi_t}}, \quad (1)$$

giving the fraction of viable networks in terms of the trace $\text{Tr}_{\mathbf{J}, \xi_t}$ over the phase space (regulatory interactions \mathbf{J} and the expression levels of transcription factors ξ_t) and an indicator function $I(\mathbf{J}, \xi)$ of couplings and all expression levels. The indicator function is defined to equal one for a viable network and zero otherwise.

Specifically, genes are labelled $i = 1, \dots, N$ for structural genes and $i = N + 1, \dots, \beta N$ for transcription factors. The regulatory network is encoded in a matrix of regulatory interactions J_{ij} , with positive J_{ij} indicating that gene $j > N$ produces a transcription factor which enhances the expression of gene $i \neq j$, and represses gene i for negative values of J_{ij} . Different external and internal states of the organism are labelled by $\mu = 1, \dots, \alpha N$. ξ_i^μ denotes the (log-)expression level of gene i in state μ , and is positive for high concentrations of the gene product and negative for low concentrations. Assuming transcription factors act independently on their target genes, the condition for a viable network is modelled as

$$\xi_i^\mu / \sqrt{N} \sum_j J_{ij} \xi_j^\mu > \kappa \quad \forall i, \mu. \quad (2)$$

Threshold condition (2) has been used extensively to model neural [11] and gene regulatory networks [9, 10, 12, 13]. The indicator function for a viable network in the partition function (1) can be written in terms of the Heaviside step-function $\Theta(x)$ as

$$\begin{aligned} I(\{J_{ij}, \xi_i^\mu\}) &= \prod_{i, \mu} \Theta \left(\frac{1}{\sqrt{N}} \xi_i^\mu \sum_j J_{ij} \xi_j^\mu - \kappa \right) \\ &= \prod_{i, \mu} \int_{\kappa}^{\infty} d\lambda \int \frac{dx}{2\pi} e^{i \sum_{i\mu} \lambda_i^\mu x_i^\mu - \frac{i}{\sqrt{N}} \sum_{ij\mu} x_i^\mu \xi_i^\mu J_{ij} \xi_j^\mu}. \end{aligned} \quad (3)$$

We constrain vectors of regulatory interactions \mathbf{J}_i and of expression levels ξ_i to lie on hyperspheres. This defines

the trace over phase space (4)

$$\text{Tr} = \prod_i \int d\mu(\mathbf{J}_i) \prod_{i>N} \int d\mu(\xi_i) \quad (4)$$

with $d\mu(\mathbf{J}_i) = \prod_{j>N} dJ_{ij} \delta((\beta-1)N - \sum_{j>N} J_{ij}^2)$ and analogously for the transcription factor expression levels. The quenched average of (1) over expression levels of structural genes $\langle \ln Z \rangle = \prod_{i \leq N} \int d\mu(\xi_i) \ln Z$ is performed using the replica trick.

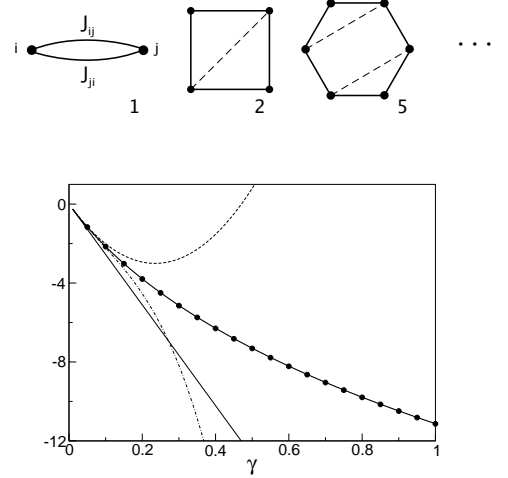


Figure 2: **Averaging over expression levels.** a) The diagrams corresponding to the first three terms in (5) are shown along with their combinatorial factors. Nodes represent variables i, j, \dots , solid lines indicate the corresponding matrix entries G_{ij} , dashed lines are contractions $i = j$. b) Plotting the logarithm of (5) against γ shows the contribution of different diagrams. The first diagram gives a linear term (thin solid line), the series up to second and third order are shown by the dashed and dashed-dotted curves respectively. These are valid approximations up to some finite values of γ only. The thick solid line gives the full series to infinite order (thick solid line) along with a numerical computation of (5), where G_{ij} was taken a random matrix of size $N = 50$ with i.i.d. normally distributed elements.

Transcription factors play a special role; their expression levels provide the regulatory input for *every gene* in the regulatory network. This produces an effective coupling between regulatory interactions of different genes. One consequence emerges already at the level of the average of (3) over the expression levels ξ . As an illustration, we consider a toy problem, where the average of $\exp\{-i\sqrt{\gamma/(2N)} \sum_{ij} \xi_i G_{ij} \xi_j\}$ is computed over a distribution of independent normally distributed variables ξ_i . G_{ij} is a symmetric matrix with uncorrelated random

entries.

$$\begin{aligned}
\langle e^{-\frac{i}{2}\sqrt{\gamma/N}\sum_{ij}\xi_i G_{ij}\xi_j} \rangle &= 1/\sqrt{\det(\mathbb{1} + i\sqrt{\gamma/N}\mathbf{G})} \\
&= \exp\left\{-\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n}\text{tr}(i\sqrt{\gamma/N}\mathbf{G})^n\right\} \\
&= \exp\left\{\gamma/4\sum_i z_i - \gamma^2/4\sum_i z_i^2 + 5\gamma^3/12\sum_i z_i^3 + \dots\right\} \\
&= \exp\left\{\sum_i w(\gamma z_i)\right\}
\end{aligned} \tag{5}$$

with shorthand $z_i = 1/N \sum_j G_{ij}G_{ji}$ and tr denoting the matrix trace. The successive terms in the power series (5) can be represented diagrammatically; Fig. 2 a) shows the first three diagrams. Fig. 2 b) shows how the different powers in (5) contribute to the average and how for finite values of γ the series has to be taken to infinite order, giving $w(z) = \frac{2z+1-\sqrt{1+4z}}{8z} - \frac{1}{2}\log\left(\frac{1}{2} + \frac{1}{2}\sqrt{1+4z}\right)$. This is in contrast to the standard situation in fully-connected disordered models, where in the thermodynamic limit the series in (5) terminates after the first term.

The approach (5) applied to the full model (1)-(3) gives $\sum_{i \leq N, \mu} w(z_i^\mu) + \sum_{i > N, \mu, a} w(z_i^{\mu a})$ with $z_i^\mu = \sum_a (x_i^{\mu a})^2$ and $z_i^{\mu a} = (x_i^{\mu a})^2 + \frac{1}{N} \sum_j (x_i^{\mu a})^2 (J_{ji}^a)^2 + \frac{2}{N} \sum_j x_i^{\mu a} x_j^{\mu a} J_{ij}^a J_{ji}^a$. Neglecting fluctuations of $z_i^{\mu a}$ across genes, the entropy of viable networks $\langle \langle S \rangle \rangle \equiv N^2 s$ can be computed in the thermodynamic limit $N \rightarrow \infty$ by standard methods. Within a replica-symmetric ansatz we obtain

$$\begin{aligned}
s &= \text{extr} \left[\frac{1}{2}\alpha(\beta-1)(F - \ln F) + \frac{1}{2}\alpha\hat{X}_2^- X_2^- \right. \\
&+ \frac{1}{2}\alpha(\beta-1)\hat{X}_2^+ X_2^+ + \alpha(\beta-1)\hat{X}_1 X_1 + \frac{(\beta-1)^2}{4} \ln(1-h^2) \\
&+ \alpha(\beta-1) \ln \left[\int_{\kappa}^{\infty} d\lambda \int \frac{dx}{2\pi} \exp \left\{ w \left(\frac{1}{F^2} [x^2 + X_2^+ \right. \right. \right. \\
&\quad \left. \left. \left. - 2ixhX_1] + \frac{X_2^-}{(\beta-1)F} \right) + i(\lambda - \hat{X}_1)x - \frac{1}{2}\hat{X}_2^+ x^2 \right\} \right] \\
&\left. + \alpha \int d\hat{y} dy e^{[i\hat{y}\hat{y} + w(\frac{y}{F})]} \ln \left[H \left(\frac{\kappa}{\sqrt{\hat{X}_2^- + 2i\hat{y}}} \right) \right] \right].
\end{aligned} \tag{6}$$

The entropy s of viable networks is determined by the extremum over the saddle point parameters $F, X_1, \hat{X}_1, X_2^\pm, \hat{X}_2^\pm$, and h . Saddle point parameter h has an intuitive interpretation in terms of the symmetry of regulatory interactions and will be discussed below. $H(x)$ denotes the cumulative Gaussian measure $\int_x^\infty \frac{dy}{\sqrt{2\pi}} \exp\{-y^2/2\}$.

The entropy of viable networks (6) decreases with increasing number of patterns $P = \alpha N$, see Fig. 3. This is to be expected, as each set of expression patterns induces a new set of constraints on the network. However, the entropy remains finite even as the number of patterns becomes large with $\alpha \rightarrow \infty$: there (typically) always

exists a viable network, and there is no transition to a phase where solutions of (2) no longer exist. Such phase transitions are well known in neural networks and many combinatorial problems [11]. In contrast, the ability of regulatory networks to store a large number of expression patterns of structural genes stems directly from the freedom to choose expression levels of transcription factors: transcription factor expression levels adapt in such a way that regulatory interactions compatible with expression levels of all genes can be found.

The saddle point parameter $h = \frac{1}{(\beta-1)^2 N^2} \sum_{i,j} J_{ij} J_{ji}$ is the *symmetry parameter* of the resulting regulatory network. A positive value of h indicates that if gene i regulates gene j , and also j regulates i , the signs of these interactions are correlated, with like signs occurring more frequently than opposite signs. The origin of this symmetry lies in condition (2) for a viable network, where a positive value of $\xi_i^\mu \xi_j^\mu$ for some i, j, μ gives rise to positive values for both J_{ij} and J_{ji} , and analogously for negative values. Thus the symmetry parameter h increases with the number of expression patterns; Fig. 3 shows the analytical result for h along with the outcome of numerical simulations.

This statistical bias towards symmetric interactions is compatible with empirical data on regulatory networks. A literature search for well-documented cases of mutually interacting genes with known interaction sign finds 9 cases of mutually interacting gene pairs with like interaction sign [14] compared to only 3 cases with different sign [15]. A nontrivial statistics of reciprocal interactions has also been found in neural and metabolic networks [16], where, however, the signs of the interactions are generally unknown.

Over long evolutionary timescales, the required expression levels of structural genes can change. In the case of enzymes, for instance, changing nutrient availability or changing metabolic rates alter the required expression levels. Such changes of the expression levels of structural genes induce adaptive changes both of the regulatory network, and of the expression levels of transcription factors. To investigate the adaptation to changing expression levels of structural genes, we systematically perturb the expression levels of structural genes of a viable network, rendering it, in general, at first unviable. (Expression levels $\xi_{i \leq N}^\mu$ are perturbed by adding i.i.d. Gaussian random variables with mean zero and standard deviation η and normalizing their variances to one again.) Subsequently, regulatory interactions and transcription factor expression levels are adapted until the viability condition (2) is satisfied again. The overlap $q_\xi^< = \frac{1}{NP} \sum_{i \leq N, \mu} \xi_i^\mu \xi_i^{\prime \mu}$ of structural gene expression levels of the unperturbed (unprimed) and the perturbed (primed) system quantifies the strength of the perturbation, the analogously defined $q_\xi^>, q_J^<, q_J^>$ quantify the response of the system to this perturbation. Figure 4 shows the overlaps as a function of perturbation strength. One finds that already small

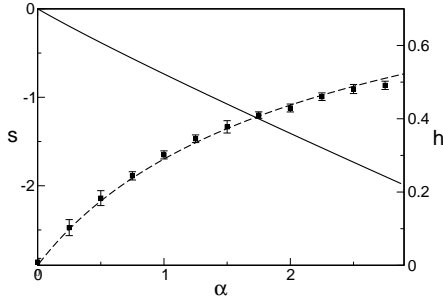


Figure 3: **Entropy and symmetry of viable networks.** With increasing number of patterns $P = \alpha N$, the space of viable networks shrinks, and the networks become increasingly symmetric, see text. Here the entropy s per structural gene (solid line) and symmetry parameter h (dashed line) are plotted against α for $\beta = 2, \kappa = 0$. The \blacksquare -symbols stem from numerical simulations with $N = 80$, averaged over 20 realizations of the quenched disorder (mean and standard error). The numerics is based on simulated annealing under Monte-Carlo dynamics of the regulatory interactions J_{ij} and the expression levels of transcription factors $\xi_{i>N}^\mu$.

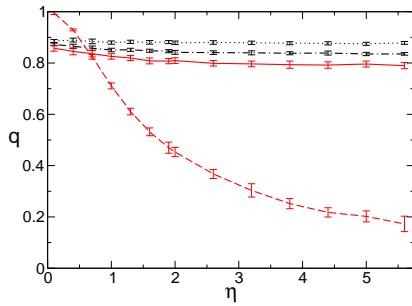


Figure 4: **Response to changing expression levels.** The overlaps of perturbed and unperturbed systems (see text) are plotted against the perturbation strength η : structural genes expression level overlaps (red dashed line) tend to zero with increasing η by construction, whereas TF expression level overlaps (red solid line) quickly reach a plateau. The same holds for regulatory interactions to TF (black dotted line), and interactions to structural genes (black dash-dotted). The plateau value decreases with the fraction $1 - \beta$ of TF in the genome. The data stem from Monte-Carlo simulations with $N = 40$, $\alpha = 1$, $\beta = 2$, and $\kappa = 0$, averaged over 20 samples.

perturbations with $q_\xi^< \approx 1$ result in a drop of the overlaps to a plateau value. Larger perturbations, and even the limit $q_\xi^< \rightarrow 0$ induce only a slow decay of $q_\xi^>, q_J^<, q_J^>$ from their plateau values. Accordingly, close to any viable network for one set of expression levels of structural genes, there exists a viable network for any other, even unrelated set of expression levels. This effect allows fast adaptation to changes in the required expression levels.

Another consequence of the observed drop of the TF expression level overlap to a plateau is that expression levels of TF change *more* than those of structural genes for small perturbations. For large perturbations, the expression levels of TF change *less* than those of structural genes. This effect may explain an apparent contradiction in the cross-species comparison of experimentally measured expression levels. A comparison of humans with other primates shows large changes of TF expression levels [17] compared to structural genes, different *Drosophila* species show only small changes of TF expression levels compared to structural genes [18].

In summary, we have investigated the degeneracy of regulatory networks within a simple model of genetic regulation. Whereas the connection between annealed TF expression levels and the large space of viable networks is likely to persist also in more complex models, the geometry of this space may well change. In particular, models taking into account physical interactions between transcription factors to implement logical functions [13] lead to p-spin interactions $J_{ijk\dots}$ and may result in a disconnected solution space and combinatorial complexity.

Many thanks to A. Altland and M. Cosentino Lagomarsino for discussions. Funding from the DFG is acknowledged under grant BE 2478/2-1 and SFB 680.

-
- [1] A. Tsong, M. Miller, R. Raisner, and A. Johnson, *Cell* **115**, 389 (2003).
 - [2] R. Schröder, *Nature* **422**, 621 (2003).
 - [3] A. Tanay, A. Regev, and R. Shamir, *Proc. Natl. Acad. Sci. USA* (2005).
 - [4] D. Tautz, *Current Opinion in Genetics & Development* **10**, 575 (2000).
 - [5] V. Mustonen and M. Lässig, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15936 (2005).
 - [6] T. Chouard, *Nature* **456**, 300 (2008).
 - [7] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker, *Proc Biol Sci* **255**, 279 (1994).
 - [8] S. Bornholdt and K. Sneppen, *Phys. Rev. Lett.* **81**, 236 (1998).
 - [9] S. Ciliberti, O. Martin, and A. Wagner, *PLoS Comput. Biol.* **3**, e15 (2007).
 - [10] S. Ciliberti, O. Martin, and A. Wagner, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 13591 (2007).
 - [11] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the theory of neural computation* (AddisonWesley, Reading, MA, 1991).
 - [12] M. Wahde and J. Hertz, *Biosystems* **55**, 126 (2000).
 - [13] N. Buchler, U. Gerland, and T. Hwa, *Proc. Natl. Acad. Sci. USA* **100**, 5136 (2003).
 - [14] G. Goudreau et al., *Proc Natl Acad Sci U S A* **99**, 8719 (2002). T. Mikeladze-Dvali et al., *Cell* **122**, 775 (2005). H. C. Olguin et al., *J Cell Biol* **177**, 769 (2007). J.-L. Chew et al., *Mol Cell Biol* **25**, 6031 (2005). D. Feldser et al., *Cancer Res* **59**, 3915 (1999). S. Lemeille, J. Geiselman, and A. Latifi, *BMC Microbiol* **5**, 18 (2005). A. To et al., *Plant Cell* **18**, 1642 (2006). R. Pan-

- chanathan, H. Xin, and D. Choubey, *J Immunol* **180**, 5927 (2008). L. Canaple et al., *Mol Endocrinol* **20**, 1715 (2006).
- [15] D. Alabadi et al., *Science* **293**, 880 (2001). J. E. Schmidt, G. von Dassow, and D. Kimelman, *Development* **122**, 1711 (1996). S. Borrelli et al., *BMC Mol Biol* **8**, 85 (2007).
- [16] D. Garlaschelli and M. I. Loffredo, *Phys. Rev. Lett.* **93**, 268701 (2004).
- [17] Y. Gilad, A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White, *Nature* **440**, 242 (2006).
- [18] S. A. Rifkin, J. Kim, and K. P. White, *Nat Genet* **33**, 138 (2003).